Income Prediction

محمد ابراهيم جمال الدين ابراهيم 2021170445
مهند خالد علي محمد 2021170558
بيشوي سدرة صابر سدرة 2021170130
مازن سامح عبد الفتاح صابر 2021170437
مصطفى محمود مصطفى يونس 2021170531

- Preprocessing Techniques
- Visualization and Analysis
- Models and Hyperparameters used

Preprocessing Techniques

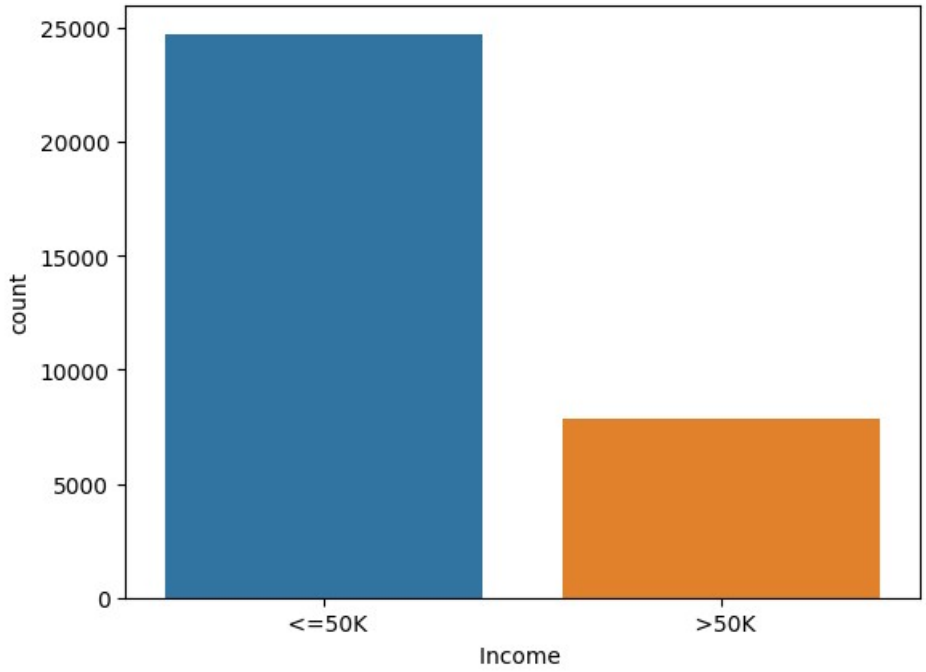
- **Mode Imputation**
  - We decided to replace null values with their mode as we had no more than 5% of the data missing (null values) .

- **Dealing with outliers**
  - We replaced any value more than (q3+1.5*IQR) with (q3+1.5*IQR)
  - We replaced any value less than (q1-1.5*IQR) with (q1-1.5*IQR)

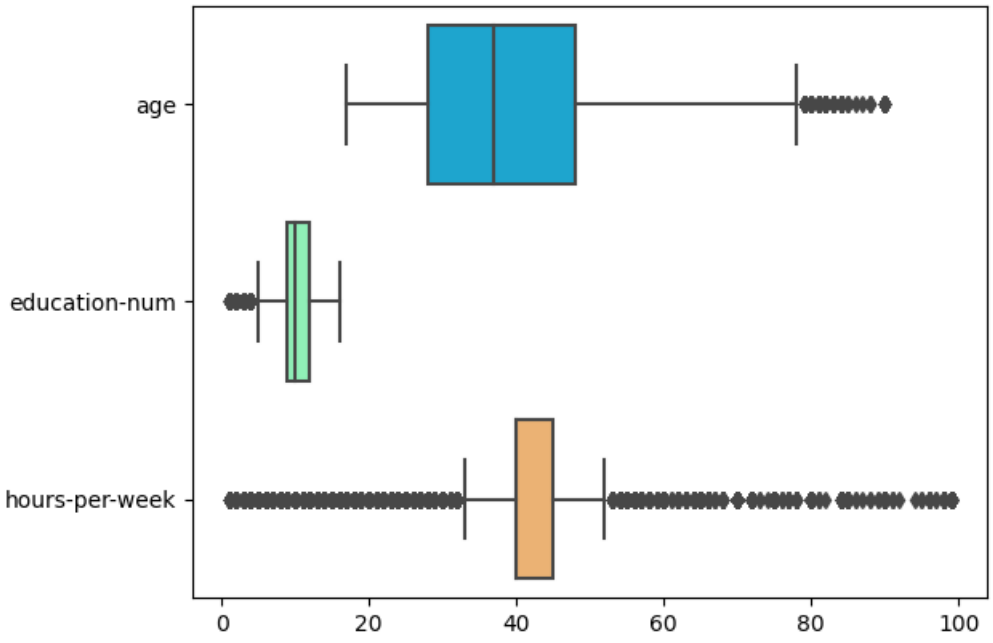
- **Feature Selection**
  - We decided to drop :
    - “capital-gain” , “capital-loss” , “race” : biased data
    - “relationship” : had the same effect as “marital status”
    - “fnlwgt” : not relevant to our target “income”

Visualization

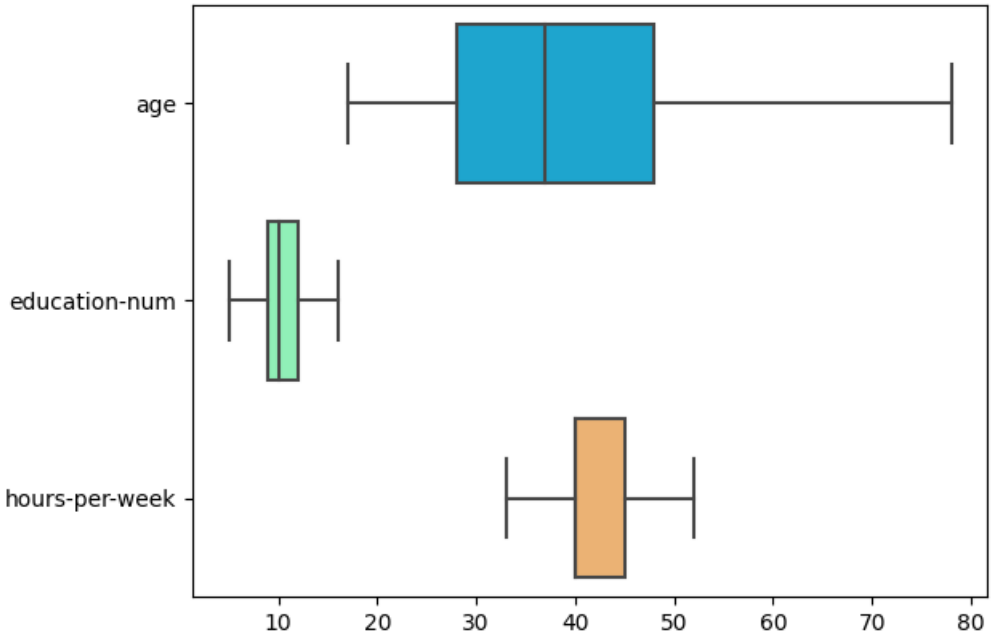
- Income count



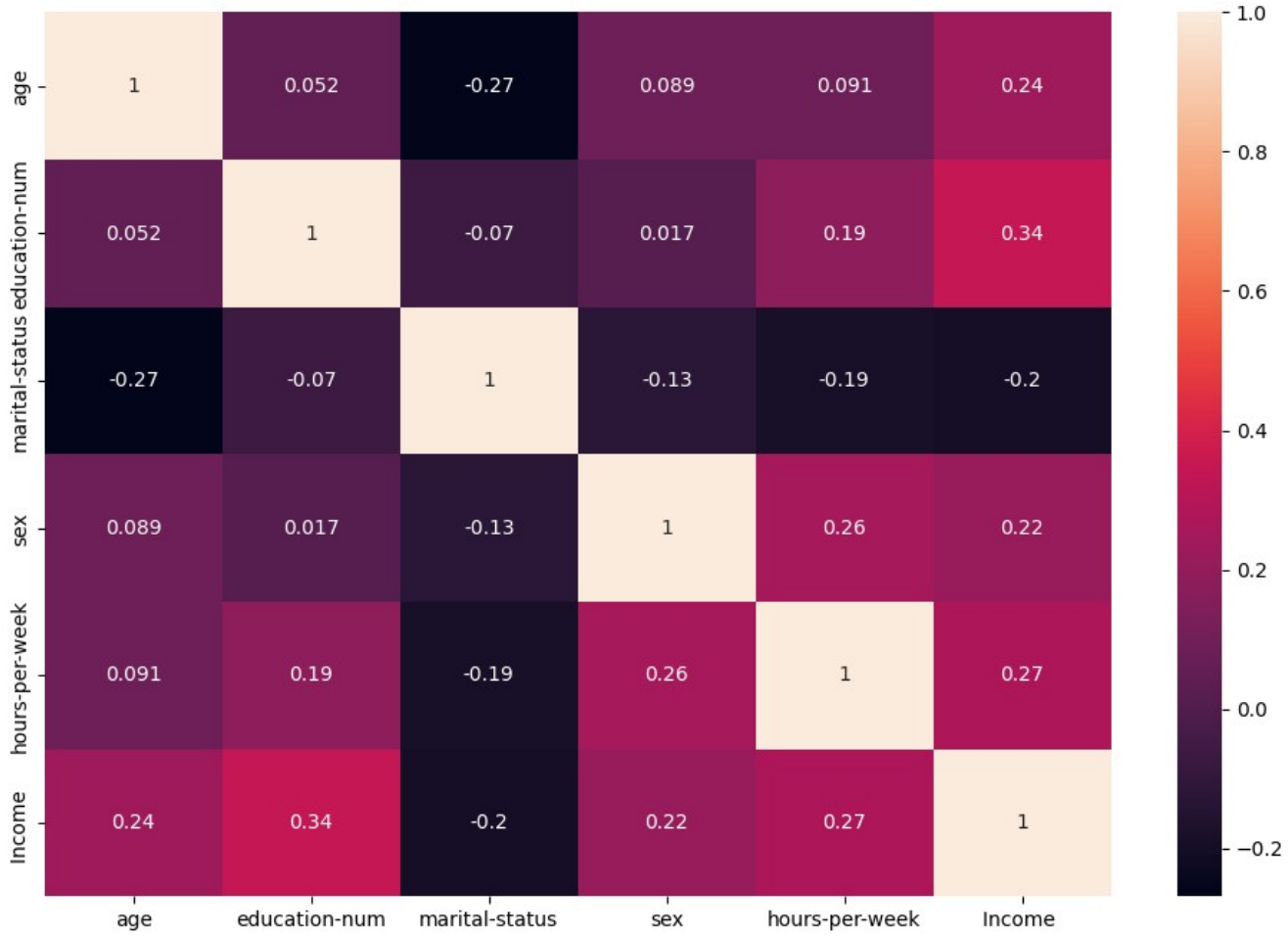
- Boxplot BEFORE handling the outliers



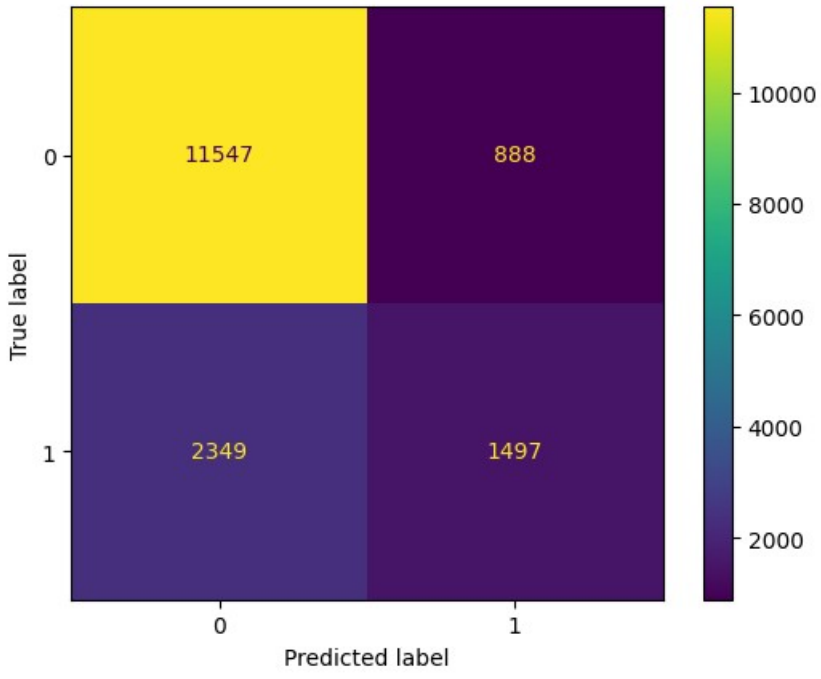
- Boxplot AFTER handling the outliers



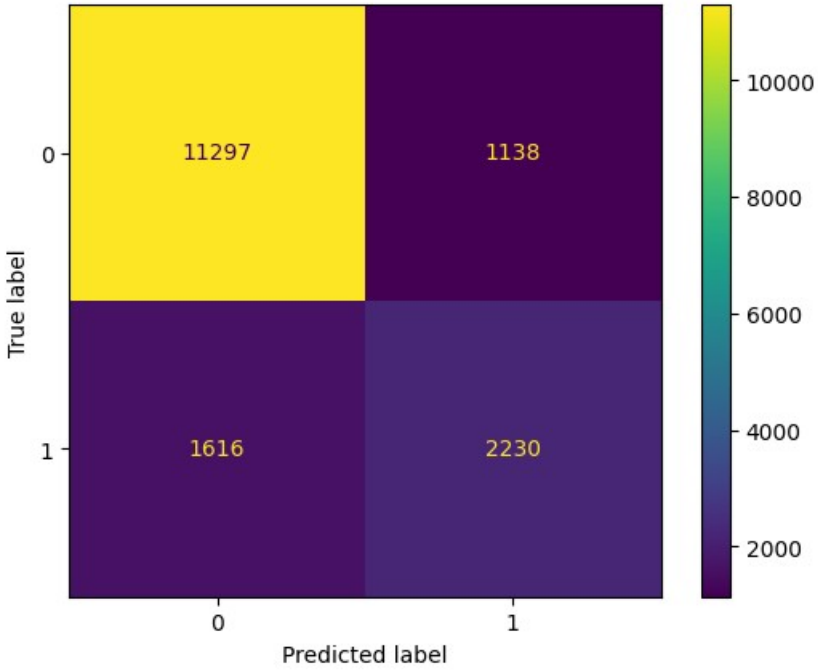
- Correlation



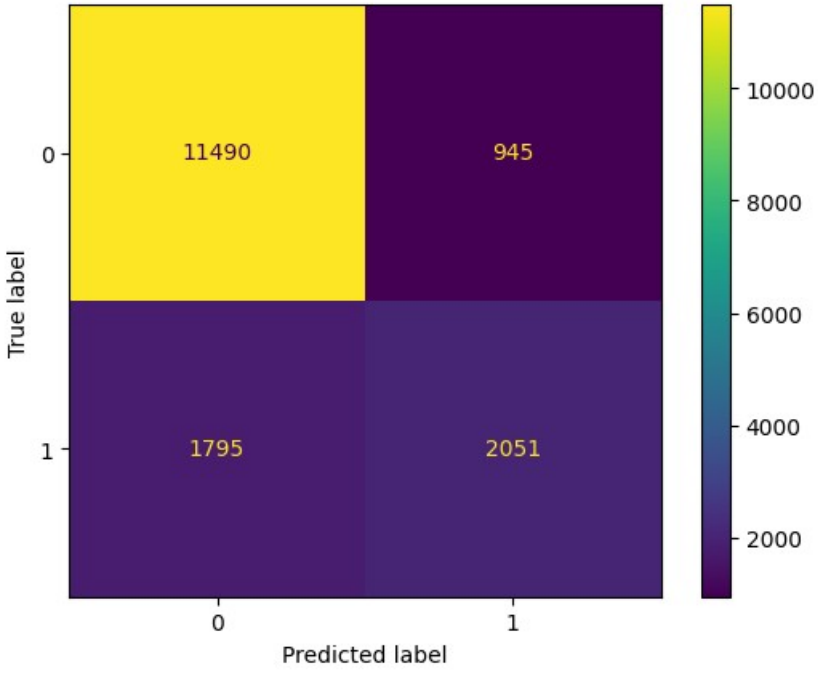
- Logistic Regression (Confusion Matrix)



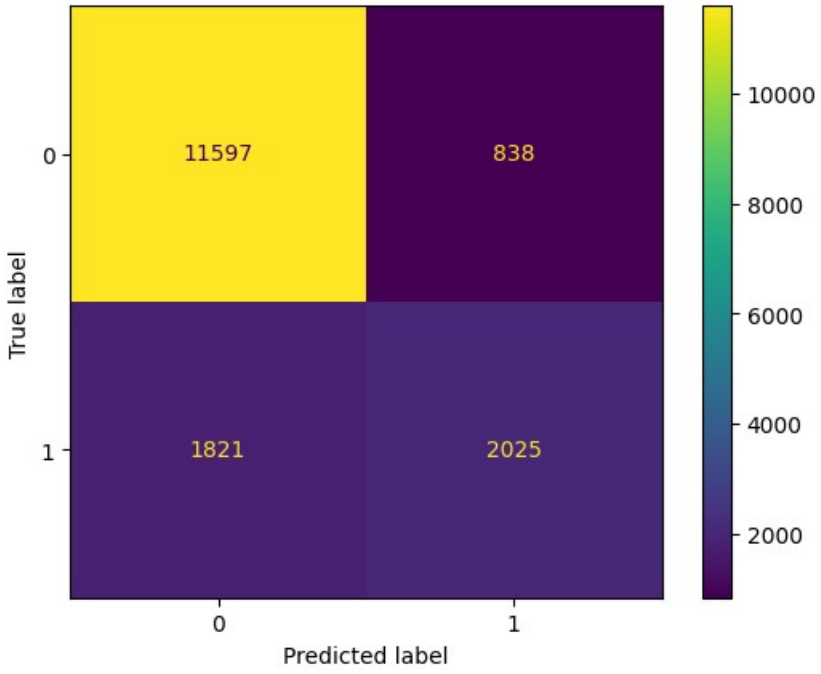
- Decision Tree (Confusion Matrix)



- SVM (Confusion Matrix)



- RandomForest(Confusion Matrix)



Analysis

- We studied the dataset and found that there were many categorical features so we converted them to numerical features by LabelEncoding
- Due to the wide range of our features' values we had to scale all the features using StandardScaler
- We found out that there were no nulls but instead there were (?) mark so we replaced all "?" With Nulls
- Correlation was weak between all features so we abandoned this method in our feature selection

Models and Hyperparameters

- **Logistic Regression : (accuracy 80%)**
  - **Solver Function is the main hyperparameter here .. we tried tuning it but the default(*lbfgs*) seemed to have the best accuracy**
- **Decision Tree : (accuracy 83%)**
  - **We used Grid Search Function and tuned the hyperparameters .. this led to increasing the accuracy from 82% when using the default to 83%**
- **SVM : (accuracy 83%)**
  - **Kernel Function is the main hyperparameter here .. we tried tuning it as well but the default(*rbf*) seemed to have the best accuracy**
- **Random Forest : (accuracy 84%)**
  - **We tried tuning two hyperparametrs .. the Max-Depth of each forest(different decision tree) to 10 and N-estimators to 100**

Other Techniques

- **Grid Search(decision tree hyperparameter tuning) :**
  - We use it as it is a hyperparameter optimization technique used to find the best combination of hyperparameters for a machine learning model. It involves exhaustively searching through a predefined grid of hyperparameter values and evaluating the model's performance on each combination.

Conclusion

After tuning the different hyperparameters in each model and trying out 4 models (**Logistic Regression , Decision Tree , SVM , Random Forest**) .. We set the default hyperparameters in Logistic Regression and SVM but tuned the Decision Tree and Random Forest.

We concluded that Random Forest got the best accuracy in the 4 models with 84% accuracy